CSC2547 Presentation: Curiosity-driven exploration

Count-based VS Info gain-based

Sheng Jia, Tinglin Duan
(First year master students)

Planning to Be Surprised: Optimal Bayesian Exploration in Dynamic Environments

Yi Sun, Faustino Gomez, and Jürgen Schmidhuber

VIME: Variational Information Maximizing Exploration

Rein Houthooft[§†‡], **Xi Chen**[†‡], **Yan Duan**[†‡], **John Schulman**[†‡], **Filip De Turck**[§], **Pieter Abbeel**[†‡]

Unifying Count-Based Exploration and Intrinsic Motivation

Marc G. Bellemare
bellemare@google.com

Sriram Srinivasan
srsrinivasan@google.com

Georg Ostrovski
ostrovski@google.com

Tom Schaul
schaul@google.com

David Saxton
saxton@google.com

Rémi Munos
munos@google.com

1. PLAN (2011)

1. VIME (NeurIPS2016)

1. CTS (NeurIPS2016)

Outline

- Motivation, Related Works and Demo
- Planning to Be Surprised
- Variational Information Maximizing Exploration
- Unifying Count-Based Exploration and Intrinsic Motivation
- Comparisons and Discussion

Outline

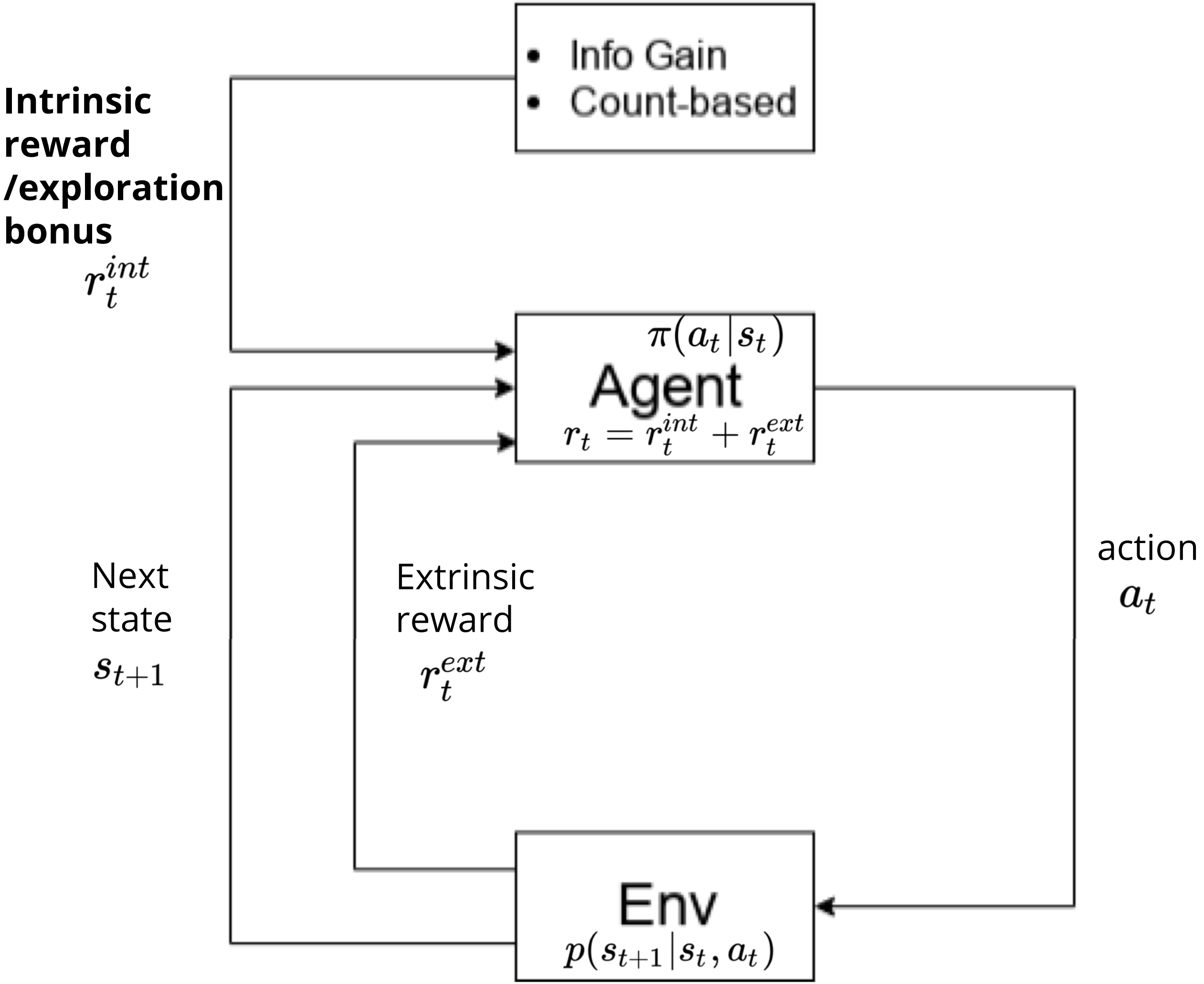
- Motivation, Related Works and Demo
- Planning to Be Surprised
- Variational Information Maximizing Exploration
- Unifying Count-Based Exploration and Intrinsic Motivation
- Comparisons and Discussion

Background

RL+Curiosity

History:

$$\xi_t = (s_1, a_1, s_2, a_2, \ldots, s_t)$$



What is exploration?

Intrinsic motivation:

- Reducing the agent's uncertainty over the environment's dynamics.

[Plan] $p\left(s_{t+1}|\xi_t, a_t; \theta\right)$

$p(\theta)$

[VIME] $p\left(s_{t+1}|s_t, a_t; \theta\right)$

[CTS]

Count-based

- Use (pseudo) visitation counts to guide agents to unvisited states.

Why exploration useful?

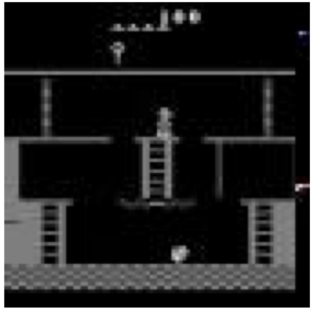
DEMO

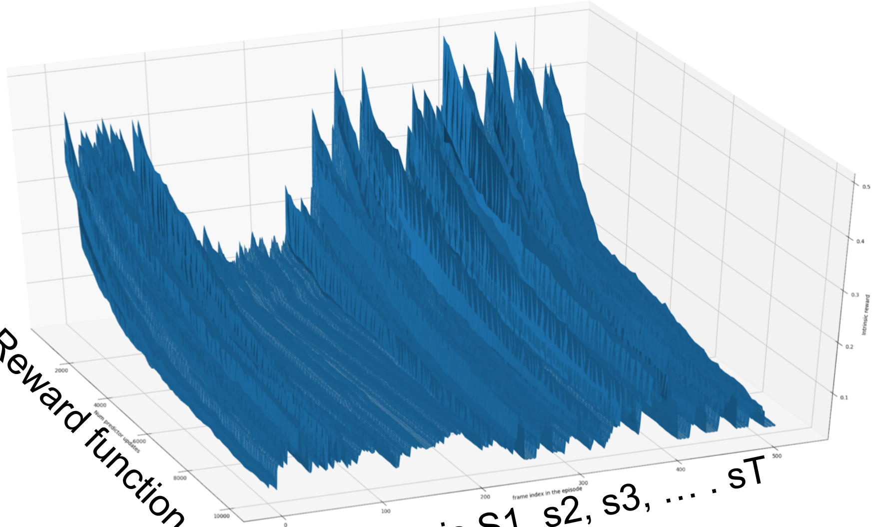
Sparse Reward Problem
Montezuma's revenge

Our original plot & demo

DQN

DQN + Exploration bonus





Related work (Timeline)

The notion of Intrinsic Motivation

2010 Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010)

2011 PLAN

Bayesian Optimal Exploration

L2 prediction error using neural networks

2015 Incentivizing Exploration In Reinforcement Learning With Deep Predictive Models

2016 VIME

Approximate "PLAN"

CTS

Pseudocount exploration

Pseudocount + Pixel CNN

2017 Count-Based Exploration with Neural Density Models

2018 Exploration by Random Network Distillation

Distillation error as a quantification of uncertainty

Pseudocount in 2016 still achieves SOTA for Montezuma's revenge"

2019 On Bonus Based Exploration Methods In The Arcade Learning Environment

Outline

- Motivation, Related Works and Demo
- Planning to Be Surprised
- Variational Information Maximizing Exploration
- Unifying Count-Based Exploration and Intrinsic Motivation
- Comparisons and Discussion

[PLAN] contribution

Dynamics model

$$p\left(s_{t+1}|\xi_t, a_t; \theta\right)$$

Bayes update for posterior distribution of the dynamics model

$$p\left(\underbrace{\theta}_{y}|\xi_t, a_t, \underbrace{s_{t+1}}_{x}\right) = \frac{p\left(\underbrace{\theta}_{y}|\xi_t, a_t\right) p\left(\underbrace{s_{t+1}}_{x}|\xi_t, a_t; \underbrace{\theta}_{y}\right)}{p\left(\underbrace{s_{t+1}}_{x}|\xi_t, a_t\right)}$$

Optimal Bayesian Exploration based on:

$$q^{\tau}\left(\xi_t, a_t\right) = \mathbb{E}_{s_{t+1}|\xi_t, a_t}\left[D_{KL}\left[p\left(\theta|\xi_t, a_t, s_{t+1}\right)||p\left(\theta|\xi_t\right)\right]\right] + \mathbb{E}_{s_{t+1}|\xi_t, a_t}\left[\max_{a_{t+1}} q_{\pi}^{\tau-1}\left(\left(\xi_t, a_t, s_{t+1}\right), a_{t+1}\right)\right]$$

Expected cumulative info gain fo tau steps if performing this action

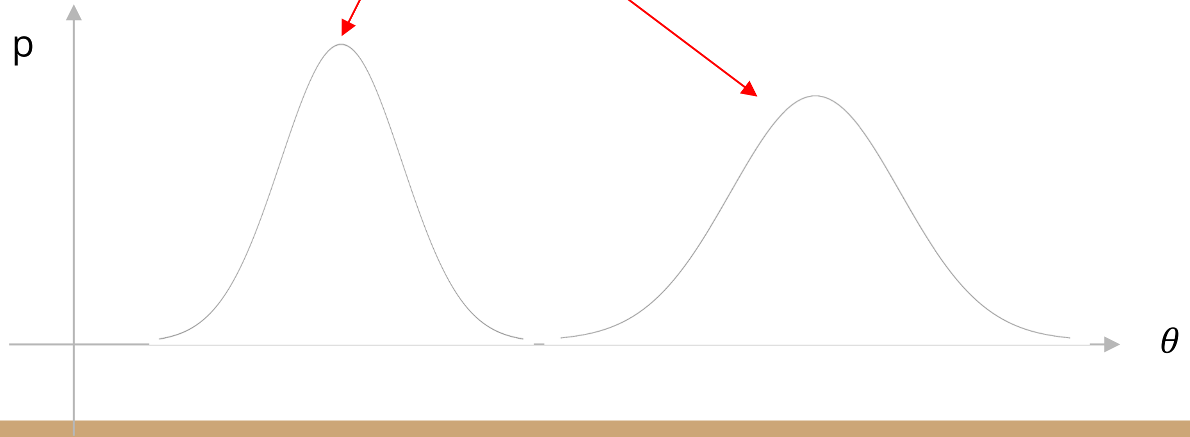
Expected one-step info gain

Expected cumulative info gain for tau-1 steps if performing this next action

[PLAN] Quantify "surprise" with info gain

$$\xi_t = (s_1, a_1, s_2, a_2, \ldots, s_t)$$
$$\xi'_t = (s_1, a_1, s_2, a_2, \ldots, s_t, a_t \ldots, s_{t'})$$

$$KL(p(\theta|\xi'_t)||p(\theta|\xi_t)) = \int p(\theta|\xi'_t) \log \frac{p(\theta|\xi'_t)}{p(\theta|\xi_t)} d\theta$$



[PLAN] 1-step expected information gain

"1-step expected info gain" "expected immediate info gain"

$$
\begin{aligned}
&\mathbb{E}_{s_{t+1}\sim P(\cdot|\xi_t,a_t)}\left[D_{KL}\left[p\left(\theta|\xi_t,a_t,s_{t+1}\right)||p\left(\theta|\xi_t\right)\right]\right] \\
&=\sum_{s_{t+1}} p\left(s_{t+1}|\xi_t,a_t\right)\int p\left(\theta|\xi_t,a_t,s_{t+1}\right)\log\frac{p\left(\theta|\xi_t,a_t,s_{t+1}\right)}{p\left(\theta|\xi_t\right)}d\theta \\
&=\sum_{s_{t+1}}\int p\left(s_{t+1},\theta|\xi_t,a_t\right)\log\frac{p\left(s_{t+1},\theta|\xi_t,a_t\right)}{\underbrace{p\left(\theta|\xi_t\right)}_{p(\theta|\xi_t,a_t)}p\left(s_{t+1}|\xi_t,a_t\right)}d\theta \\
&=I\left(S_{t+1};\Theta|\xi_t,a_t\right)
\end{aligned}
$$

NOTE: VIME uses this as the Intrinsic reward!

"Mutual info between next state distribution & model parameter"

[PLAN] “Planning to be surprised”

Cumulative τsteps info gain

Curious Q-value

$$q_{\pi}^{\tau}\left(\xi_t, a_t\right) = \mathbb{E}_{s_{t+1}, a_{t+1}, s_{t+2}, \cdots a_{t+\tau}, s_{t+\tau+1} | \xi_t, a_t} D_{KL}\left[p\left(\theta | \xi_t, a_t, s_{t+1}, a_{t+1}, \cdots, a_{t+\tau}, s_{t+\tau+1}\right) || p\left(\theta | \xi_t\right)\right]$$

Follow a policy

“Planning tau steps” because not actually observed yet

Perform an action

[PLAN] Optimal Bayesian Exploration **policy**

[Method1] Computing optimal curiosity-Q backwards for tau steps

$$q^{\tau}\left(\xi_t, a_t\right) = \mathbb{E}_{s_{t+1}|\xi_t, a_t}\left[D_{KL}\left[p\left(\theta|\xi_t, a_t, s_{t+1}\right)||p\left(\theta|\xi_t\right)\right]\right] + \mathbb{E}_{s_{t+1}|\xi_t, a_t}\left[\max_{a_{t+1}} q^{\tau-1}\left(\left(\xi_t, a_t, s_{t+1}\right), a_{t+1}\right)\right]$$

[Method2] Policy Iteration

Repeat applying

Policy evaluation

$$v_{\pi}^{\tau}\left(\xi_t\right) = \mathbb{E}_{a_t|\xi_t}\left[\mathbb{E}_{s_{t+1}|\xi_t, a_t}\left[D_{KL}\left[p\left(\theta|\xi_t, a_t, s_{t+1}\right)||p\left(\theta|\xi_t\right)\right] + v_{\pi}^{\tau-1}\left(\xi_t, a_t, s_{t+1}\right)\right]\right]$$

Policy improvement

$$\pi^{\tau}\left(\xi_t\right) = \arg\max q^{\tau}\left(\xi_t, a_t\right)$$

[Plan] Non-triviality of curious Q-value

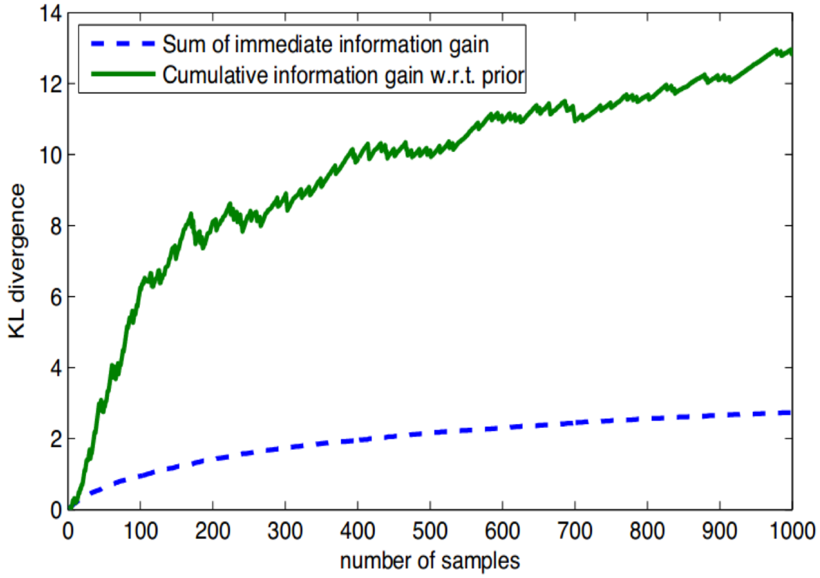
Info gain additive in expectation!

$$\mathbb{E}_{\xi_t''||\xi_t}\left[D_{KL}\left[p\left(\theta|\xi_t''\right)||p\left(\theta|\xi_t\right)\right]\right] = D_{KL}\left[p\left(\theta|\xi_t'\right)||p\left(\theta|\xi_t\right)\right] + \mathbb{E}_{\xi_t''||\xi_t}\left[D_{KL}\left[p\left(\theta|\xi_t''\right)||p\left(\theta|\xi_t'\right)\right]\right]$$

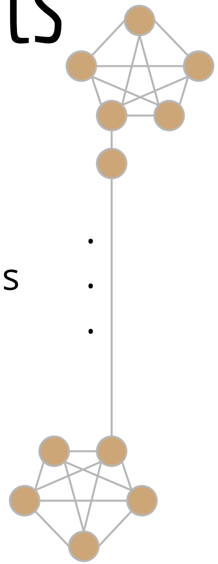
Cumulative != Sum

$$D_{KL}\left[p\left(\theta|\xi_t''\right)||p\left(\theta|\xi_t\right)\right] \neq D_{KL}\left[p\left(\theta|\xi_t'\right)||p\left(\theta|\xi_t\right)\right] + D_{KL}\left[p\left(\theta|\xi_t''\right)||p\left(\theta|\xi_t'\right)\right]$$

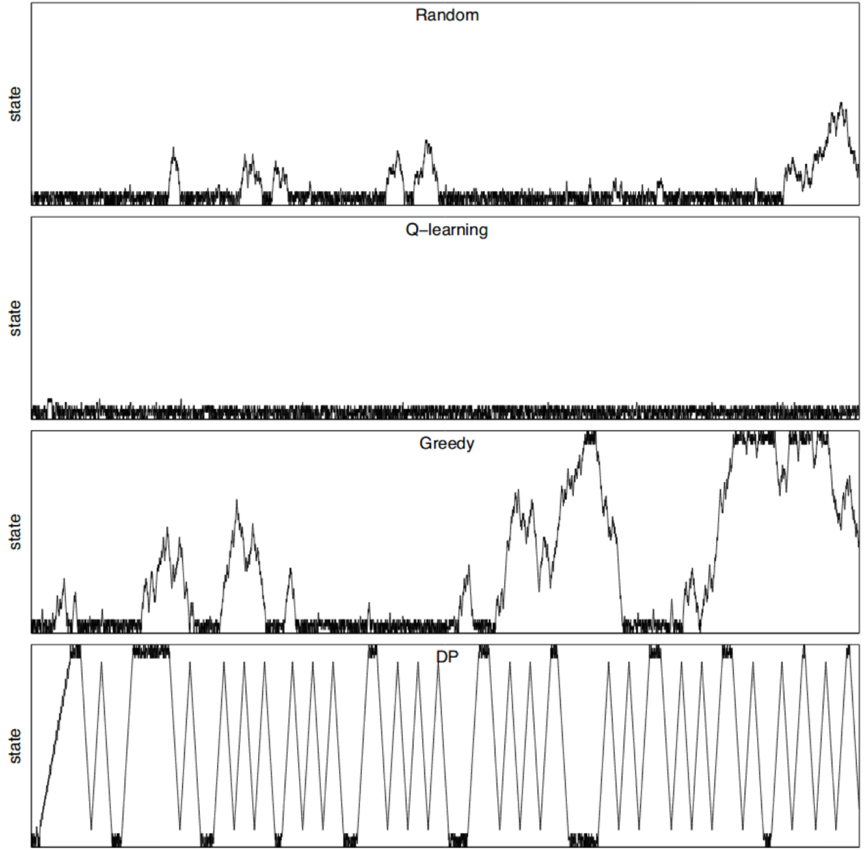
Cumulative information gain fluctuates!



[Plan] Results



50 states



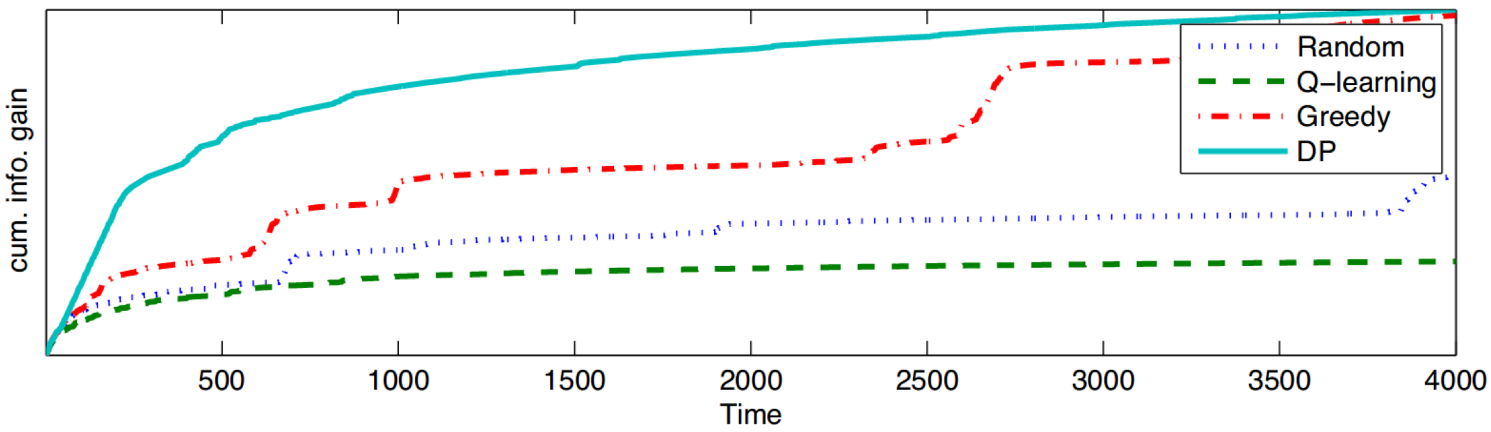
Random

Greedy w.r.t expected one-step info gain

Q-learning using one-step info gain

Policy iteration (Dynamic programming approximation to optimal bayesian exploration)

[Plan] Results



Outline

- Motivation, Related Works and Demo
- Planning to Be Surprised
- **Variational Information Maximizing Exploration**
- Unifying Count-Based Exploration and Intrinsic Motivation
- Comparisons and Discussion

[VIME] contribution

Dynamic
s model

$$p\left(s_{t+1} \middle| s_t, a_t; \theta\right)$$

Variational inference for posterior distribution of dynamics model

$$\phi' = \arg\min_{\phi} \Big[\overbrace{\underbrace{D_{\mathrm{KL}}[q(\theta;\phi) \,\|\, q(\theta;\phi_{t-1})]}_{\ell_{\mathrm{KL}}(q(\theta;\phi))} - \mathbb{E}_{\theta\sim q(\cdot;\phi)} \left[\log p(s_t|\xi_t, a_t;\theta)\right]}^{\ell(q(\theta;\phi), s_t)} \Big]$$

1-step exploration bonus

$$r'(s_t, a_t, s_{t+1}) \leftarrow r(s_t, a_t) + \eta D_{\mathrm{KL}}[q(\theta;\phi'_{n+1}) \,\|\, q(\theta;\phi_{n+1})]$$

[VIME] Quantify the information gained

$$\xi_t = (s_1, a_1, s_2, a_2, \ldots, s_t)$$

Reminder: PLAN cumulative info gain

$$KL(p(\theta|\xi_t')||p(\theta|\xi_t))$$

$$D_{\mathrm{KL}}[p(\theta|\xi_t, a_t, s_{t+1}) \,\|\, p(\theta|\xi_t)]$$

[VIME] Variational Bayes

What's hard?

$$p\left(\underbrace{\theta}_{y}|\xi_t, a_t, \underbrace{s_{t+1}}_{x}\right) = \frac{p\left(\underbrace{\theta}_{y}|\xi_t, a_t\right) p\left(\underbrace{s_{t+1}}_{x}|\xi_t, a_t; \underbrace{\theta}_{y}\right)}{p\left(\underbrace{s_{t+1}}_{x}|\xi_t, a_t\right)}$$

Computing posterior for highly parameterized models (e.g. neural networks)

Approximate posterior $q(\theta;\phi) \approx p(\theta|\xi_t, a_t, s_{t+1})$ by minimizing $D_{KL}\left[q(\theta;\phi)\,||p(\theta|\xi_t, a_t, s_{t+1})\right]$

Minimize negative ELBO

$$\phi' = \arg\min_{\phi}\left[\overbrace{\underbrace{D_{\mathrm{KL}}[q(\theta;\phi)\,\|\,q(\theta;\phi_{t-1})]}_{\ell_{\mathrm{KL}}(q(\theta;\phi))} - \mathbb{E}_{\theta\sim q(\cdot;\phi)}\left[\log p(s_t|\xi_t, a_t;\theta)\right]}^{\ell(q(\theta;\phi), s_t)}\right]$$

[VIME] Optimization for variational bayes

How to minimize negative ELBO? $\ell(q(\theta;\phi), s_t)$

Take an efficient **single second-order (Newton) update step** to minimize negative ELBO:

$$\Delta\phi = H^{-1}(\ell)\nabla_\phi \ell(q(\theta;\phi), s_t)$$

[VIME] Estimate 1-step expected info gain

What's hard? $\mathbb{E}_{s_{t+1}\sim\mathcal{P}(\cdot|\xi_t,a_t)}\left[D_{\mathrm{KL}}[p(\theta|\xi_t,a_t,s_{t+1})\,\|\,p(\theta|\xi_t)]\right]$

Computing the exact one-step expected info-gain. High-dimensional states

→ Monte-carlo estimation.

$$a_t \sim \pi_\alpha(s_t) \qquad s_{t+1} \sim \mathcal{P}(\cdot|s_t,a_t)$$

$$r'(s_t,a_t,s_{t+1}) = r(s_t,a_t) + \eta D_{\mathrm{KL}}[p(\theta|\xi_t,a_t,s_{t+1})\,\|\,p(\theta|\xi_t)]$$

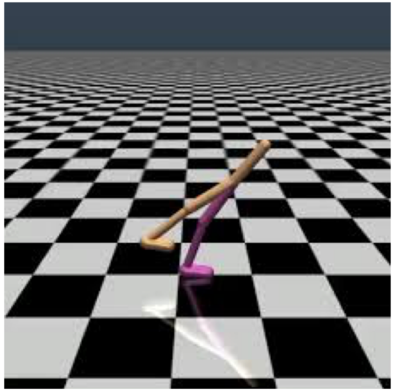
[VIME] Results (Walker-2D)

Dense reward

$$\mathcal{S} \subseteq \mathbb{R}^{20}$$

$$\mathcal{A} \subseteq \mathbb{R}^{6}$$

RL algorithm: TRPO



Average extrinsic return

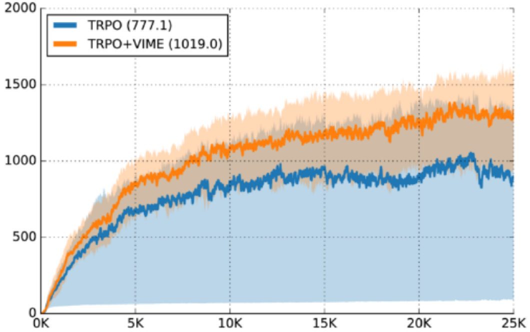


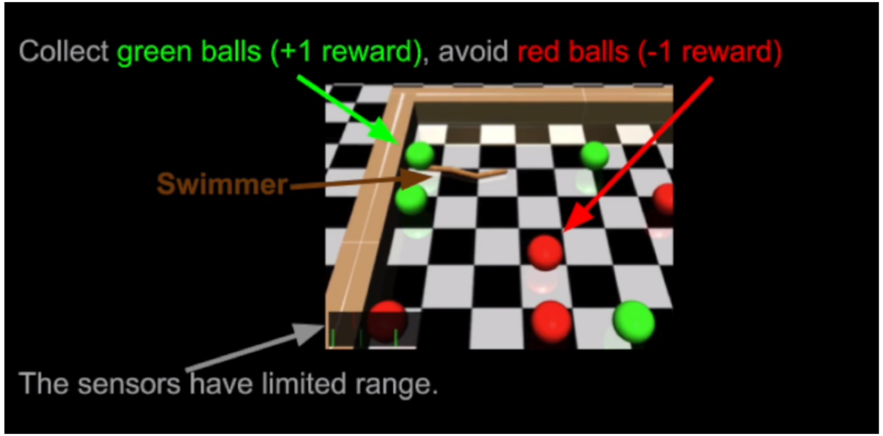
Figure 3: Performance of TRPO with and without VIME on the high-dimensional Walker2D locomotion task.

[VIME] Results (Swimmer-Gather)

Sparse reward

$$\mathcal{S} \subseteq \mathbb{R}^{33}$$
$$\mathcal{A} \subseteq \mathbb{R}^{2}$$

RL algorithm: TRPO



Average extrinsic return

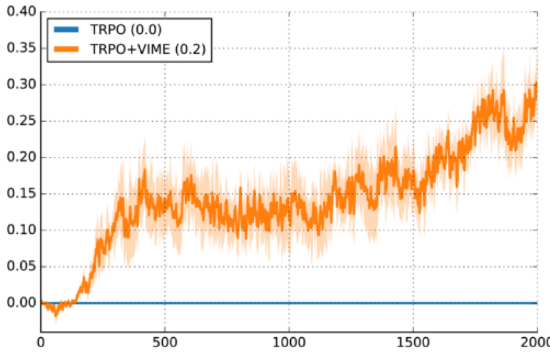


Figure 5: Performance of TRPO with and without VIME on the challenging hierarchical task SwimmerGather.

Outline

[CTS] contribution

States Density model

$$\rho_n(x)$$

Pseudo-count

$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)}$$

1-step exploration bonus

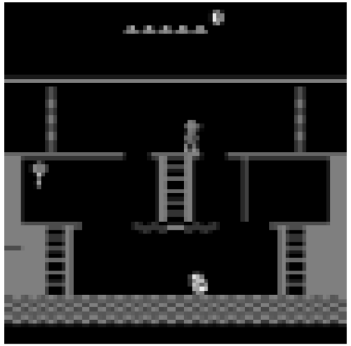
$$R_n^+(x, a) := \beta(\hat{N}_n(x) + 0.01)^{-1/2}$$

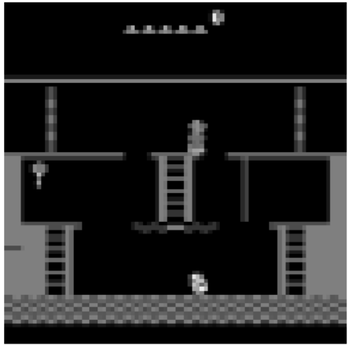
[CTS] Count state visitation

Empirical distribution

$$\mu_n(x) := \mu(x\,;\,x_{1:n}) := \frac{N_n(x)}{n}$$

Empirical count





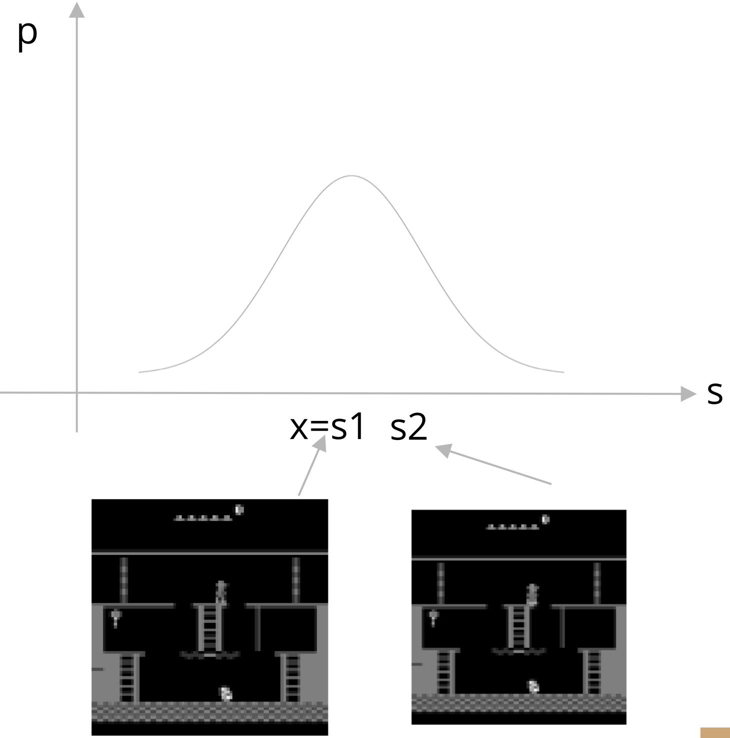
These two are different states!

But we want to increment visitation counts for both when visiting either one.

Pixel difference

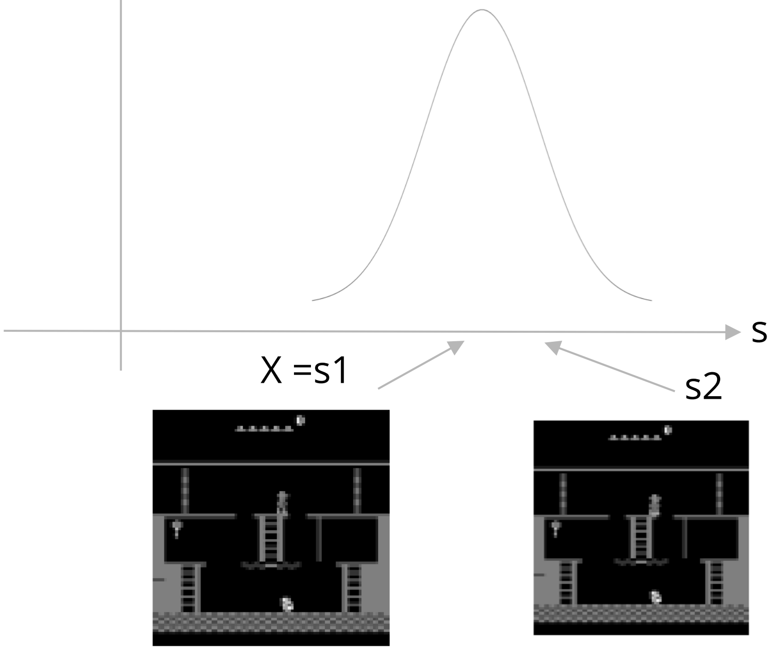
[CTS] Introduce state density model

$$\rho_n(x) := \rho(x\,;\,x_{1:n})$$



$$\rho'_n(x) := \rho(x\,;\,x_{1:n}x)$$

$$\rho'_n(x) = \Pr_\rho(X_{n+2} = x \mid X_1 \ldots X_n = x_{1:n}, X_{n+1} = x)$$



How to update CTS density model?

Check the “context tree switching” paper!
https://arxiv.org/abs/1111.3182

This was the difficulty of reading this paper as it only shows a bayes rule update for mixture of density models

$$\underbrace{\omega_n(\rho, x)}_{p(y|x)} = \frac{\overbrace{\omega_n(\rho)}^{p(y)}\,\overbrace{\rho(x; x_{1:n})}^{p(x|y)}}{\underbrace{\int_{\rho\in\mathcal{M}} \omega_n(\rho)\,\rho(x; x_{1:n})\,d\rho}_{p(x)}}$$

Remark: For pixel-cnn density model in “Count-based exploration with **neural density model**”, just **backprop**.

[CTS] Derive pseudo-count from density model

pseudo-count function $\hat{N}_n(x)$

pseudo-count total $\hat{n}$

Two constraints:
Linear system

$$\rho_n(x) = \frac{\hat{N}_n(x)}{\hat{n}} \qquad \rho'_n(x) = \frac{\hat{N}_n(x) + 1}{\hat{n} + 1}$$

Solve linear system

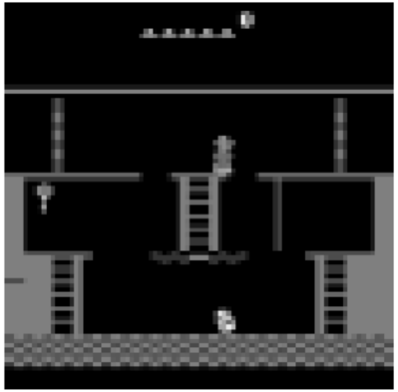
Pseudo-count derived!

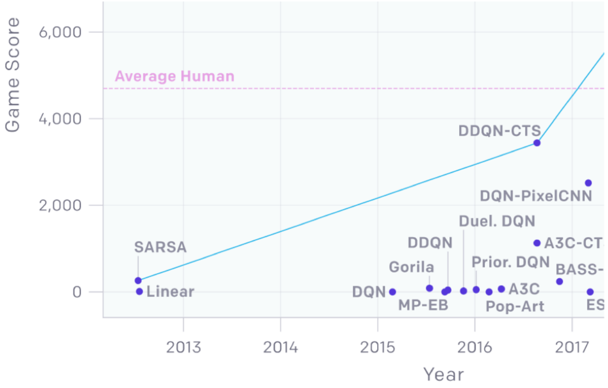
$$\hat{N}_n(x) = \frac{\rho_n(x)(1 - \rho'_n(x))}{\rho'_n(x) - \rho_n(x)}$$

[CTS] Results (Montezuma's Revenge)

State: 84x84x4
Actions: 18

RL algorithm: Double DQN





Outline

- Motivation, Related Works and Demo
- Planning to Be Surprised
- Variational Information Maximizing Exploration
- Unifying Count-Based Exploration and Intrinsic Motivation
- Summary, Comparisons and Discussion

Deriving posterior dynamics model/ density model

PLAN

Bayes rule

$$p\Big(\underbrace{\theta}_{y}|\xi_t, a_t, \underbrace{s_{t+1}}_{x}\Big) = \frac{p\Big(\underbrace{\theta}_{y}|\xi_t, a_t\Big)\, p\Big(\underbrace{s_{t+1}}_{x}|\xi_t, a_t; \underbrace{\theta}_{y}\Big)}{p\Big(\underbrace{s_{t+1}}_{x}|\xi_t, a_t\Big)}$$

VIME

Variational inference

$$\phi' = \arg\min_{\phi} \Big[\overbrace{\underbrace{D_{\mathrm{KL}}[q(\theta;\phi)\,\|\,q(\theta;\phi_{t-1})]}_{\ell_{\mathrm{KL}}(q(\theta;\phi))} - \mathbb{E}_{\theta\sim q(\cdot;\phi)}\left[\log p(s_t|\xi_t, a_t;\theta)\right]}^{\ell(q(\theta;\phi), s_t)} \Big]$$

CTS

Bayes rule

$$\underbrace{\omega_n(\rho, x)}_{p(y|x)} = \frac{\underbrace{\omega_n(\rho)}_{p(y)}\, \underbrace{\rho(x; x_{1:n})}_{p(x|y)}}{\underbrace{\int_{\rho\in\mathcal{M}} \omega_n(\rho)\, \rho(x; x_{1:n})\, d\rho}_{p(x)}}$$

Derive exploratory policy

Policy trained with the reward augmented by **intrinsic reward.**

[VIME] 1-step Information gain

$$\text{Intrinsic reward} = D_{\mathrm{KL}}[p(\theta|\xi_t, a_t, s_{t+1})\,||\,p(\theta|\xi_t)]$$

[CTS] Pseudo-count

$$\text{Intrinsic reward} = \frac{1}{\sqrt{\hat{N}_n(x)}} \qquad \hat{N}_n(x) = \frac{\rho_n(x)(1-\rho'_n(x))}{\rho'_n(x)-\rho_n(x)}$$

[PLAN] Directly argmax(curiosity Q)

Pseudo-count VS Intrinsic Motivation

Mixture model

$$\underbrace{\omega_n(\rho, x)}_{p(y|x)} = \frac{\overbrace{\omega_n(\rho)}^{p(y)}\,\overbrace{\rho(x; x_{1:n})}^{p(x|y)}}{\underbrace{\int_{\rho \in \mathcal{M}} \omega_n(\rho)\,\rho(x; x_{1:n})\,d\rho}_{p(x)}}$$

$$\mathbf{IG}_n(x) := \mathbf{IG}(x\,;\,x_{1:n}) := \mathbf{KL}\big(w_n(\cdot, x) \,\|\, w_n\big)$$

“Unifying count-based exploration and intrinsic motivations”!

$$\boldsymbol{IG}_n(x) \le \hat{N}_n(x)^{-1/2}$$

Limitations & Future Directions

PLAN → Intractable posterior & use dynamics model for expectation
Difficult to be scaled outside Tabular RL.

VIME → Currently maximize sum of 1-step info gain.

CTS → which density model leads to better generalization over states?

Learning rates of policy network VS Updating dynamic model/density model.

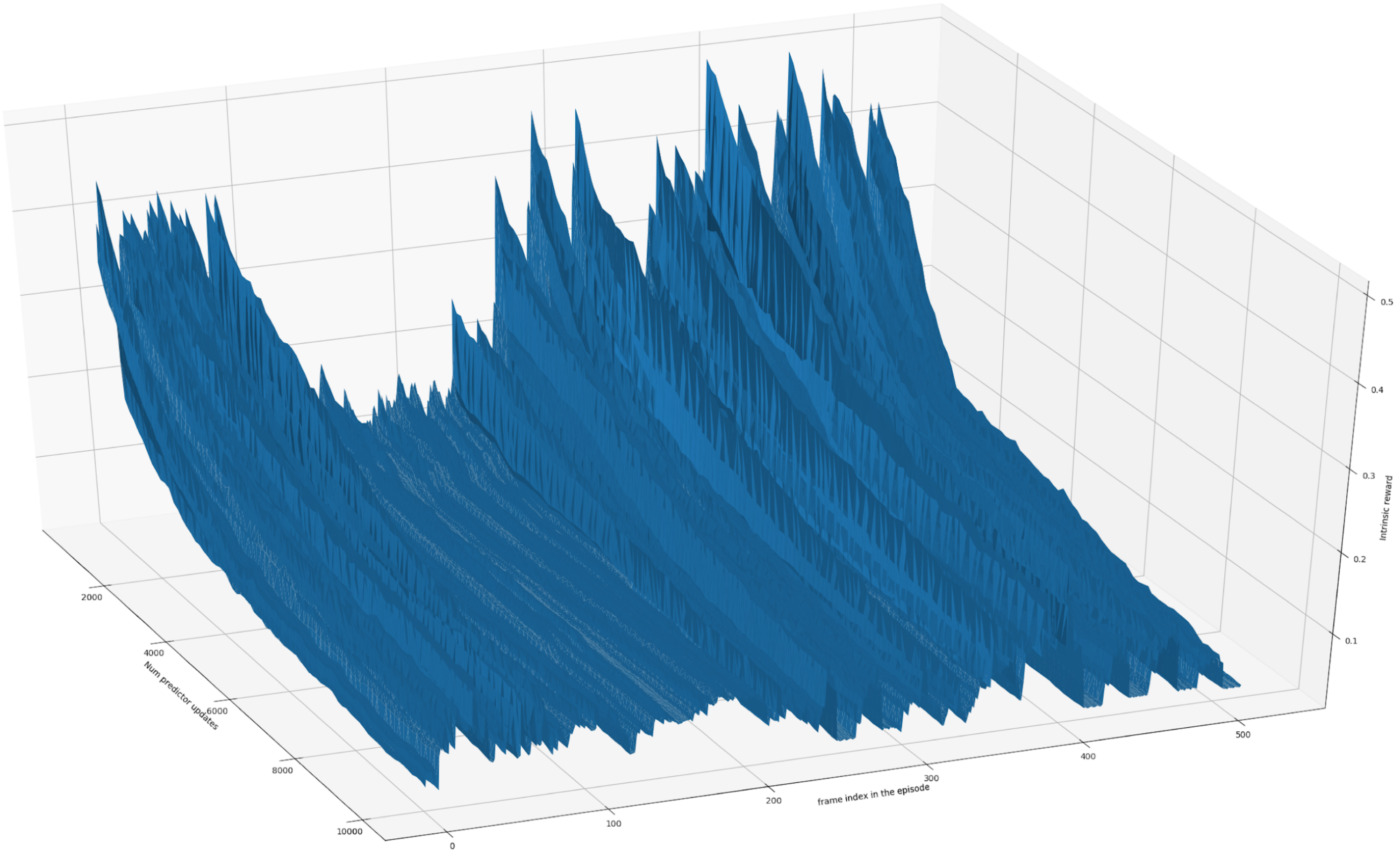
Thank you!

(Appendix)

Our derivation for "Additive in expectation"

$$
\begin{aligned}
&\mathbb{E}_{p(h''|h')}\, g\left(h''||h\right) = \mathbb{E}_{p(h''|h')}\left[KL\left(p\left(\theta|h''\right)||p\left(\theta|h\right)\right)\right] \\
&= \int\int p\left(h''|h'\right) p\left(\theta|h''\right) \log \frac{p\left(\theta|h''\right)}{p\left(\theta|h\right)} d\theta dh'' \\
&= \int\int p\left(h''|h'\right) p\left(\theta|h''\right) \log \frac{p\left(\theta|h''\right) p\left(\theta|h'\right)}{p\left(\theta|h\right) \underbrace{p\left(\theta|h'\right)}_{\text{Mult and div}}} d\theta dh'' \\
&= \int\int p\left(h''|h'\right) \underbrace{p\left(\theta|h''\right)}_{=p(\theta|h'',h')} \log \frac{p\left(\theta|h'\right)}{p\left(\theta|h\right)} d\theta dh'' + \int\int p\left(h''|h'\right) p\left(\theta|h''\right) \log \frac{p\left(\theta|h''\right)}{p\left(\theta|h'\right)} d\theta dh'' \\
&= \int\int p\left(\theta, h''|h'\right) \log \frac{p\left(\theta|h'\right)}{p\left(\theta|h\right)} d\theta dh'' + \mathbb{E}_{p(h''|h')}\, g\left(h''||h'\right) \\
&= \int \left(\int p\left(h''|\theta, h'\right) p\left(\theta|h'\right) dh''\right) \log \frac{p\left(\theta|h'\right)}{p\left(\theta|h\right)} d\theta + \mathbb{E}_{p(h''|h')}\, g\left(h''||h'\right) \\
&= \int \cancelto{1}{\left(\int p\left(h''|\theta, h'\right) dh''\right)} p\left(\theta|h'\right) \log \frac{p\left(\theta|h'\right)}{p\left(\theta|h\right)} d\theta + \mathbb{E}_{p(h''|h')}\, g\left(h''||h'\right) \\
&= KL\left(p\left(\theta|h'\right)||p\left(\theta|h\right)\right) + \mathbb{E}_{p(h''|h')}\, g\left(h''||h'\right) \\
&= g\left(h'||h\right) + \mathbb{E}_{p(h''|h')}\, g\left(h''||h'\right)
\end{aligned}
$$

h'' contains h'



0.5
0.4
0.3
0.2
0.1
Intrinsic reward
2000
4000
6000
8000
10000
Num predictor updates
0
100
200
300
400
500
frame index in the episode